Insurance Fraud Model

응용통계학과 201352099 최문석
응용통계학과 201452024 박상희
응용통계학과 201452041 이인풍

Index

Index

Feature 1

- **DANGER_PERSON**

기사에 따르면 우리가 일반적으로 생각하는 보험 사기인 고의 사고 유형 보다 일반인이 쉽게 접근할 수 있는 허위 과다 사고 유형이 훨씬 더 많은 것으로 나타났다.

또한 특정 인구학적 특성을 가진 사람이 보험사기자일 확률이 높은 것으로 나타났다. 따라서 30~50대의 직업이 회사원, 전업주부, 무직인 경우 보험사기자일 사람이 높은 것으로 판단하여 "YES", 그렇지 않은 사람은 "NO" 라는 값을 주어 변수를 생성한다.

산업 올해 상반기 보험사기 적발금액 4134억원

보험사기 적발인원 중 30~50대 연령층이 27,919명(전체의 64.8%)으로, 연령 구성비는 50대(25.6%), 40대(21.2%), 30대(18.0%)를 차지했고, 10대(청소년)의 보험사기가 전년대비 크게 증가(24.2%↑)하였으며, 60대 및 70대 이상의 고령층 보험사기는 계속적인 증가추세다.

혐의자들의 직업은 회사원(19.7%), 전업주부(10.4%),무직·일용직(9.3%) 順으로 구성비는 전년 동기와 유사했다.

허위(과다) 입원·진단 및 사고내용 조작 등의 허위·과다사고 유형은 3,130억원(75.7%)으로 전체 보험사기 유형 중 가장 큰 비중을 차지했고, 한편, 고의충돌·방화·자기재산손괴 등 고의사고 유형은 518억원(12.5%)으로 전년 동기대비 53억원(9.4%) 감소했다.

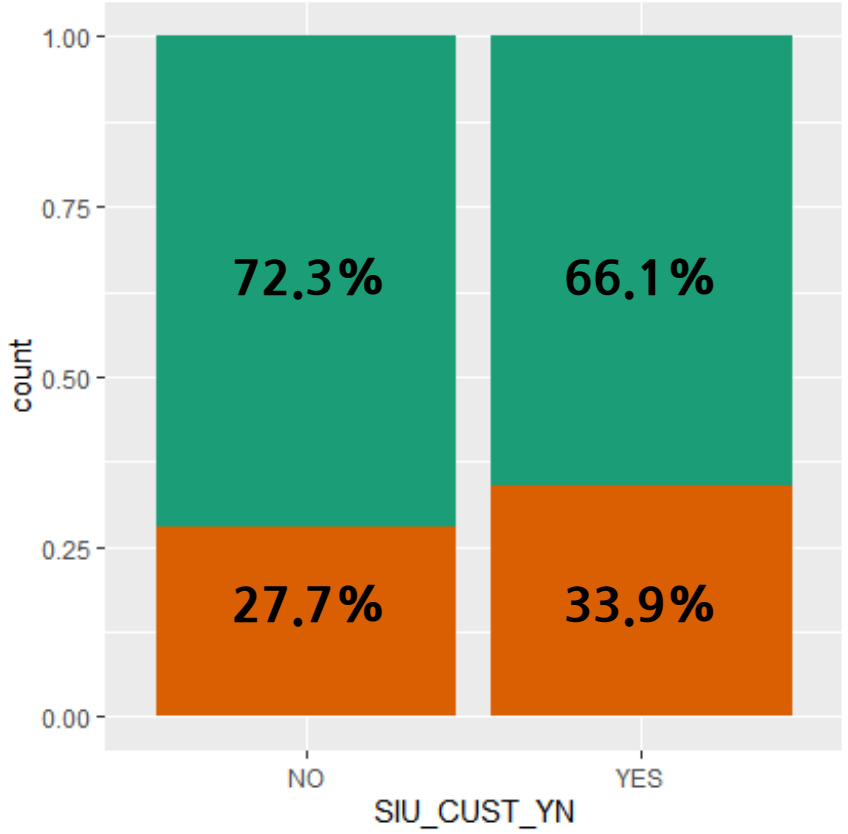
Feature 1

- **DANGER_PERSON**

실제로 보험 사기자 1,806명 중 DANGER_PERSON에 해당하는 사람은 612명으로 33.9%가 해당되었다.

DANGER_PERSON TABLE

20,607 obs		DANGER_PERSON	
		NO	YES
보험 사기	NO	13,575	5,226
	YES	1,194	612

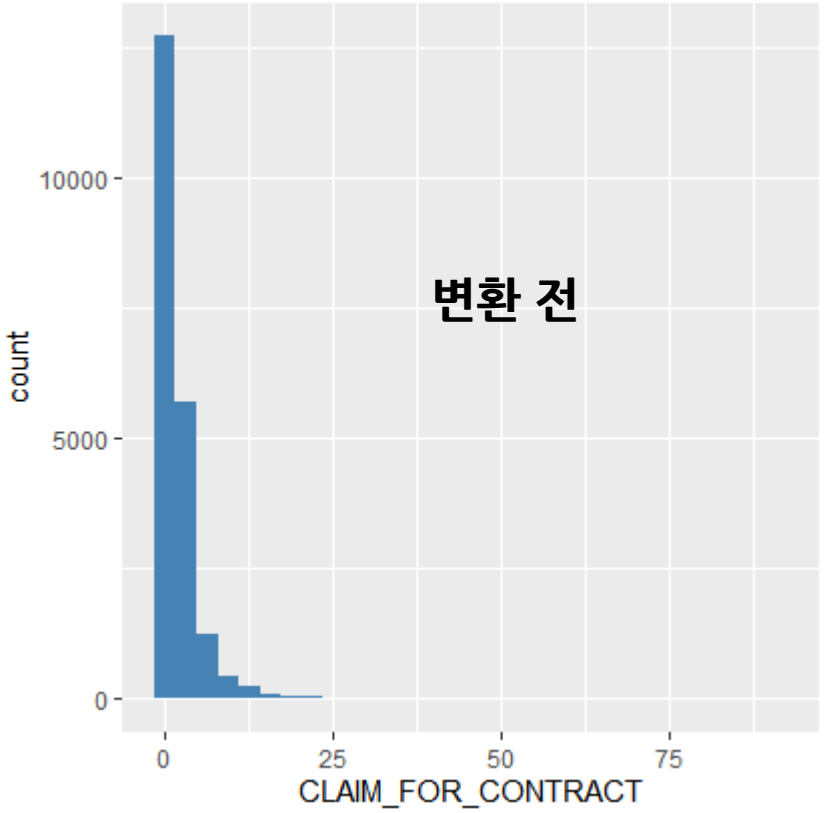


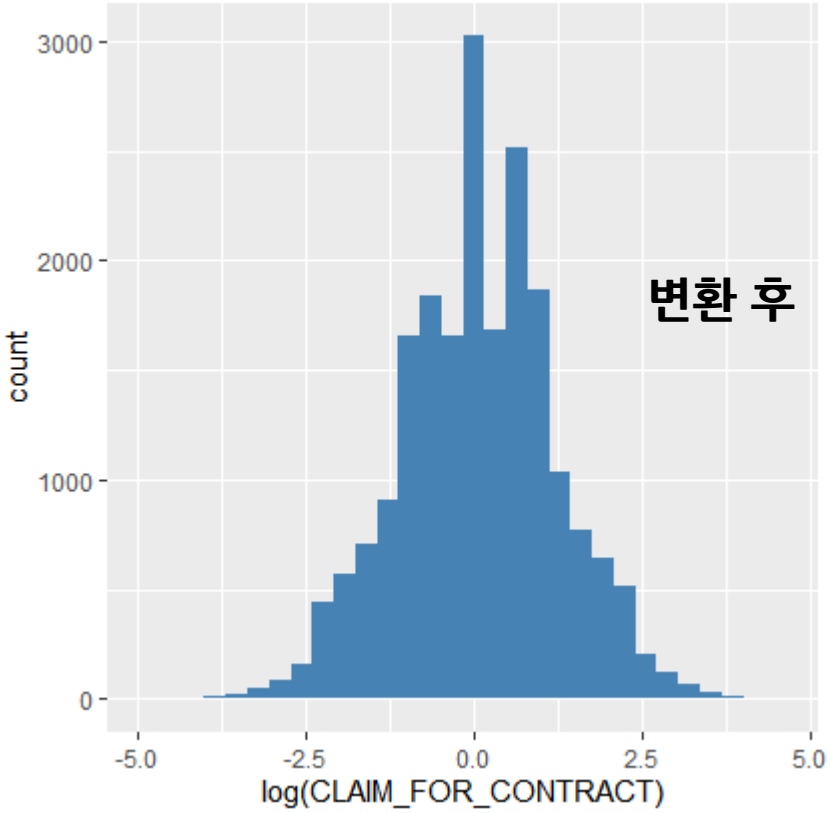
Feature 2

- **CONTRACT_FOR_CLAIM**

보험사기자일수록 계약 건수 대비 과다하게 보험을 청구할 것이라 예상하여, 고객별 계약 건수와 청구 건수를 계산한 다음 그 비율을 변수로 생성한다.

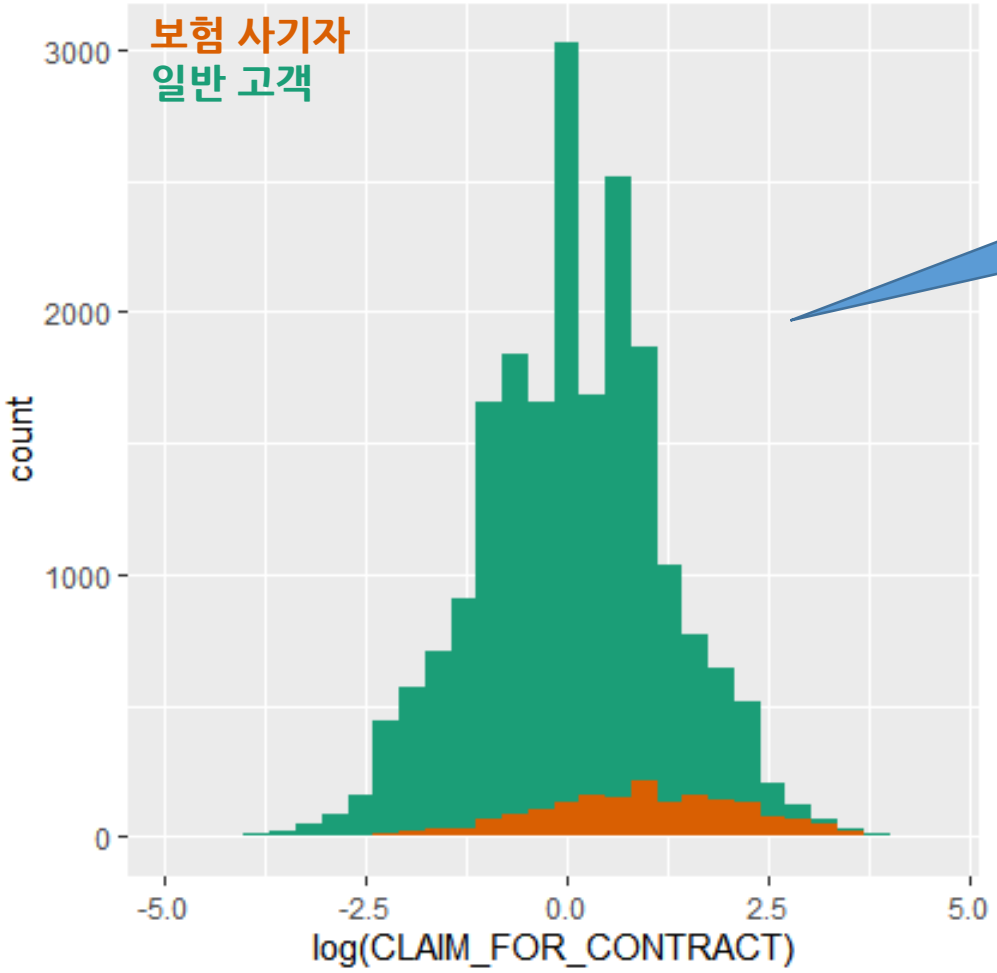
하지만, 분포를 확인해본 결과 왼쪽으로 치우쳐진 분포의 형태를 띄어 로그 변환을 진행한다.





Feature 2

- CONTRACT_FOR_CLAIM



실제로 보험사기자의 경우 분포가 오른쪽으로 좀 더 치우쳐진 것으로 나타났다.

평균 CLAIM_FOR_CONTRACT

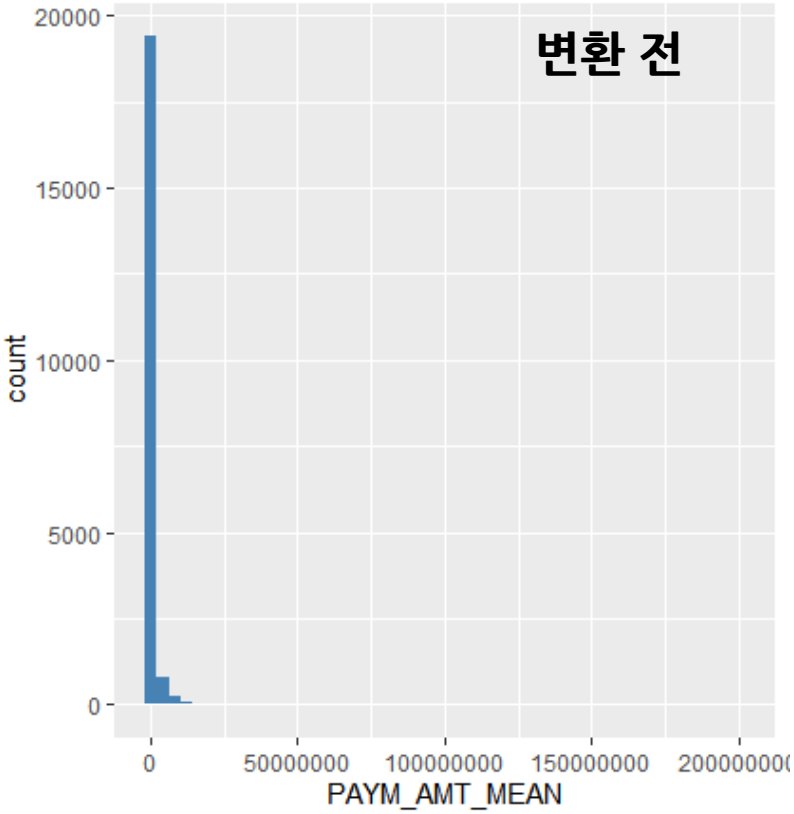
Log 변환 전	일반 고객	1.82
	보험 사기자	5.42
Log 변환 후	일반 고객	-0.004
	보험 사기자	0.932

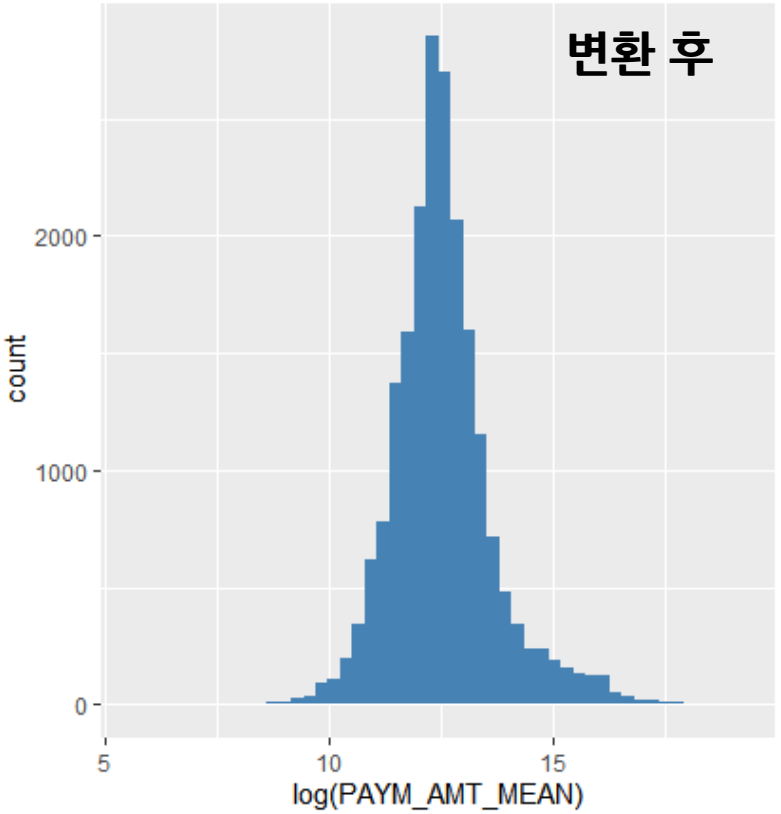
Feature 3

- **PAYM_AMT_MEAN**

보험사기자를 판별하는 데 실제로 보험 회사에서 얼마나 보험금을 지급 했는지도 중요한 변수이다.

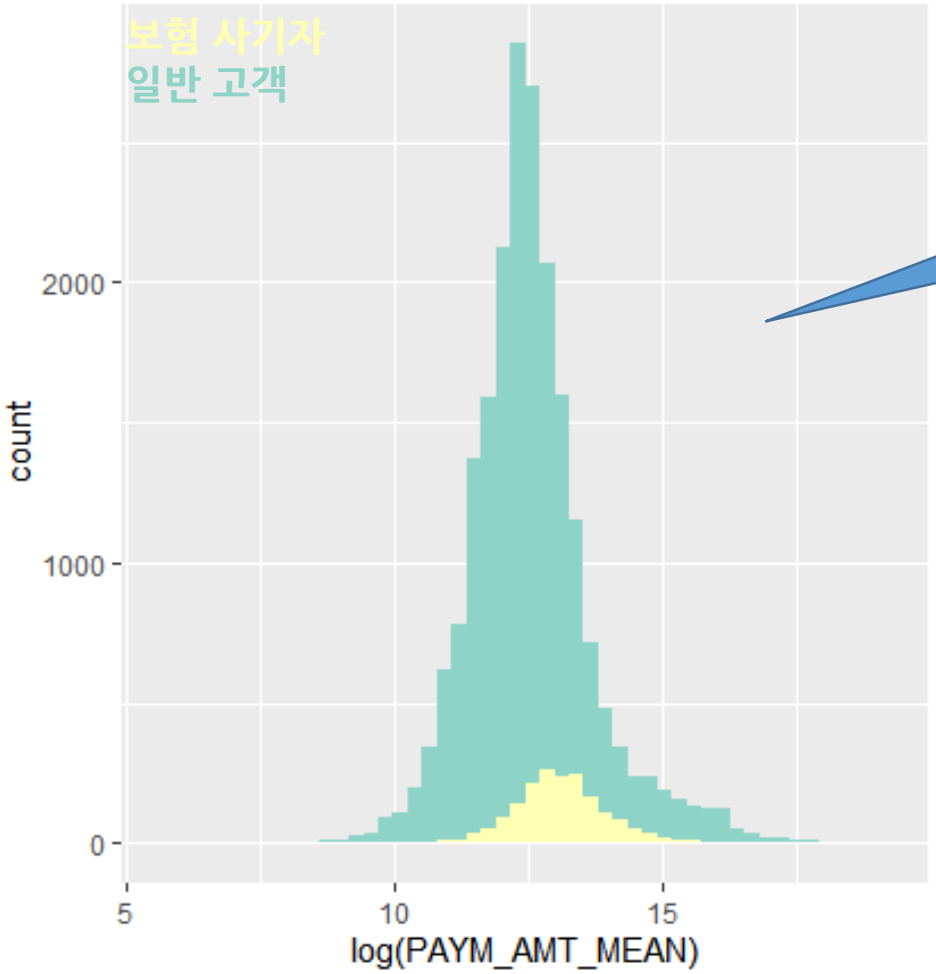
따라서, 각 고객들의 청구 건에 대한 평균 실지급액을 계산한다. 역시 로그 변환을 진행한다.





Feature 3

- PAYM_AMT_MEAN



실제로 보험사기자의 경우 분포가 오른쪽으로 좀 더 치우쳐진 것으로 나타났다.

평균 CLAIM_FOR_CONTRACT

Log 변환 전	일반 고객	810,228
	보험 사기자	898,676
Log 변환 후	일반 고객	12.5
	보험 사기자	13.2

Feature 4

- **ORIENTAL_MEDICAL**

보험사기자들은 일반적으로 사기를 칠 때 대형병원을 이용하기보다는 한방병원, 한의원 등을 주로 이용하는 것으로 나타났다.

따라서, 고객들의 보험 청구 건 중 한의원과 한방병원을 이용한 청구의 건수를 COUNT 하여 변수를 생성한다.

한의원-보험설계사 결탁 '보험사기' 72명 덜미

[뉴시스] 입력 2013.11.18 18:45

【수원=뉴시스】노수정 기자 = 한의사와 짜고 보험사기 행각을 벌인 보험설계사 등이 경찰에 무더기로 적발됐다.

경기 수원중부경찰서는 사기 등의 혐의로 김모(40)씨 등 한의사 4명을 불구속 입건했다고 18일 밝혔다.

또 이모(48·여)씨 등 보험설계사 2명과 보험가입자 66명 등 70명을 불구속 입건했다.

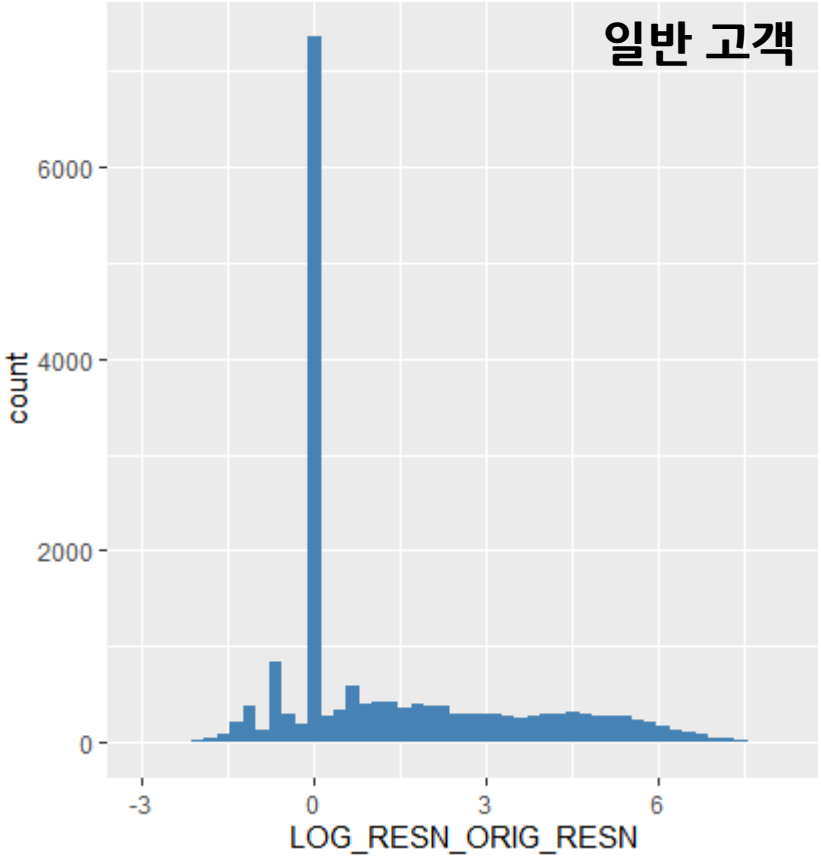
한의사들은 보약을 처방할 경우 환자 1인당 30만원의 수익이 생겨 범행에 응했다고 경찰은 전했다.

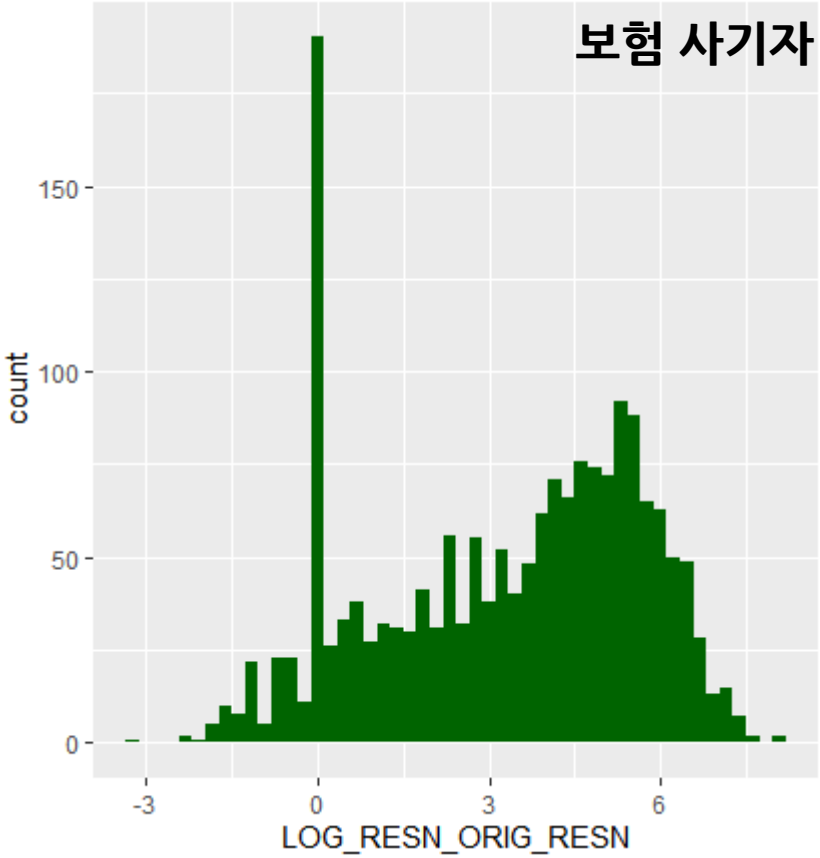
경찰 관계자는 "적발된 설계사들은 경력 20년 이상으로, 보험사에서 한의원의 경우 처방한 약이 질병치료용인지 상해치료용인지 구분할 수 없고 서류만으로 보험금을 지급한다는 점을 노렸다"고 설명했다.

Feature 5

- **RESN_ORIG_RESN**

청구 사유일자와 청구 원사유일자와의 차이(사유 - 원사유)를 계산하여 변수로 추가한다. 이 두 개의 날짜의 사이가 클수록 청구가 필요없음에도 고의적으로 보험금을 위해 청구하는 보험 사기일 가능성이 높다고 판단. 실제로 로그 변환 후의 분포를 살펴보면, 보험사기자일수록 차이가 큰 것으로 나타났다.



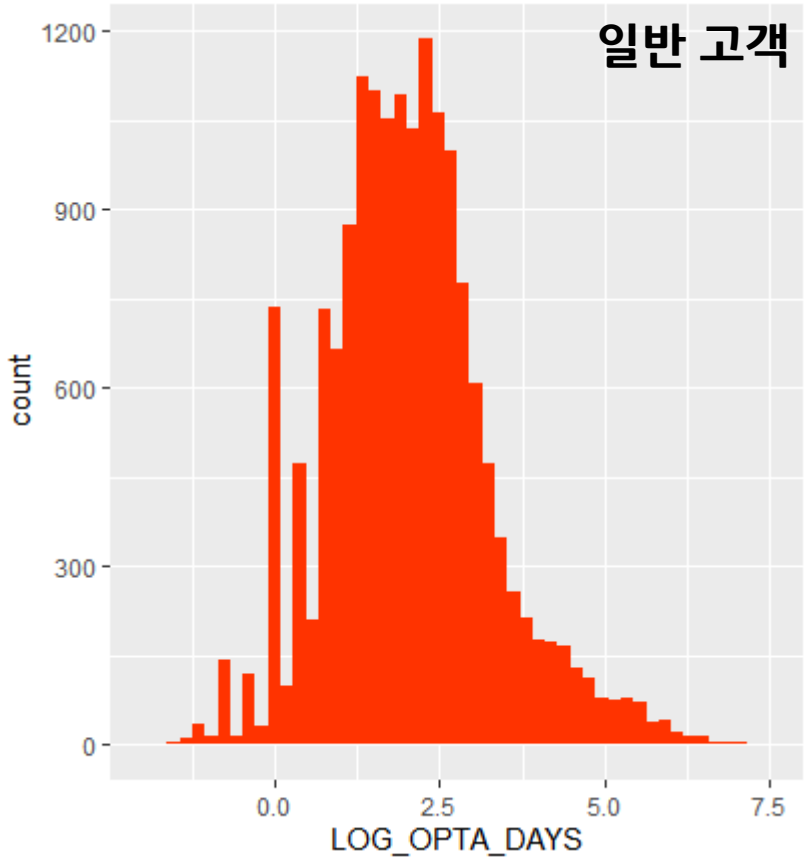


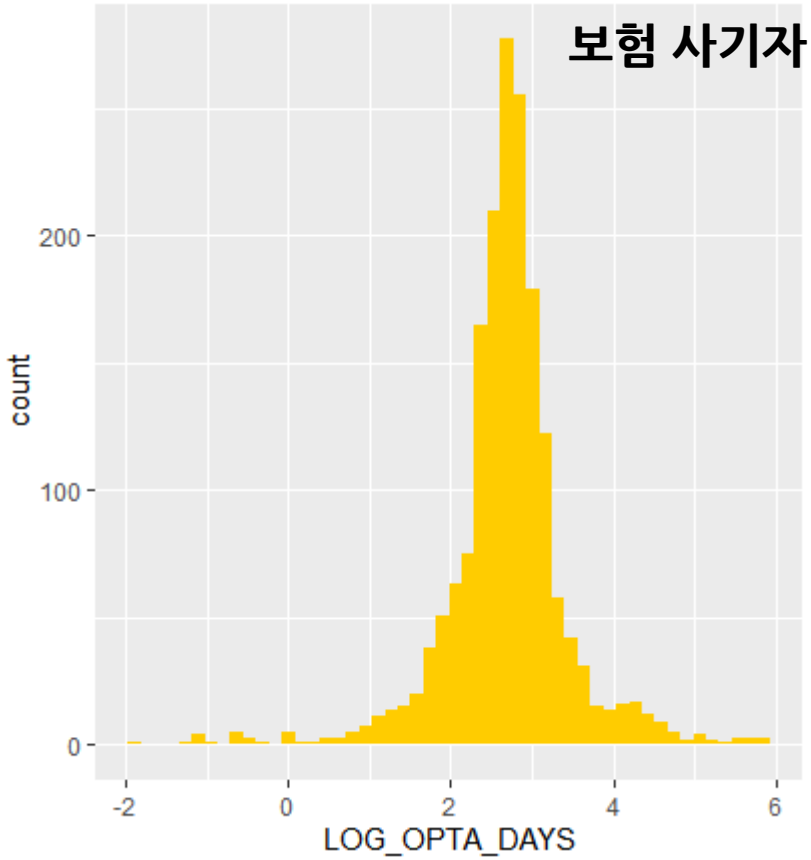
Feature 6

- **OPTA_DAYS**

  치료기간 역시 보험사기자를 판별하는 데 중요한 변수이다. 계산 후 로그 변환을 진행.

  치료기간 : OPTA_DAYS = HOSP_OTPA_ENDT - HOSP_OTPA_STDT + 1





Feature 7 ~ 9

- **ACCI_DVSN**

청구 사고 구분은 3가지 범주로 구분된다. 재해(01), 교통 재해(02), 질병(03)으로 분류된다. 각 고객들의 청구 건에 대하여 각각의 사고 구분이 몇 번 발생하였는지 COUNT 하여 변수로 추가한다.

ACCI_DVSN의 변수를 통해 ACCI_DVSN_1, ACCI_DVSN_2, ACCI_DVSN_3의 3개의 파생변수 생성

POLY_NO	CUST_ID	ACCI_DVSN
13245	1	1
5654	1	3
2798	1	2
1231	1	2
5489	2	3
222	2	1

CUST_ID	ACCI_DVSN_1	ACCI_DVSN_2	ACCI_DVSN_3
1	1	1	1
2	1	0	1
3	2	2	1

Feature 10 ~ 13

- **VLID_HOSP_OTDA_MEAN**

  평균 유효 통원/입원 일수, 각 고객들에 대하여 평균 유효 통원/입원 일수를 계산하고 로그 변환한 변수

- **COUNT_TRMT_ITEM_MEAN**

  실손 영수증 내 진료과목의 개수를 COUNT하여 로그 변환한 변수

- **NON_PAY_RATIO_MEAN**

  각 고객의 청구 건에 대한 실손비급여비율의 평균값을 로그 변환한 변수

- **DMND_CODE_2**

  지급청구의 원인이 되는 사유 코드 중 입원(02) 청구 건의 수를 COUNT하고 로그 변환한 변수

총 13개의 변수
20,607 obs

Index

Logistic Regression

Test Obs = 6,183		Pred	
		No	Yes
True	No	5,553	73
	Yes	356	201

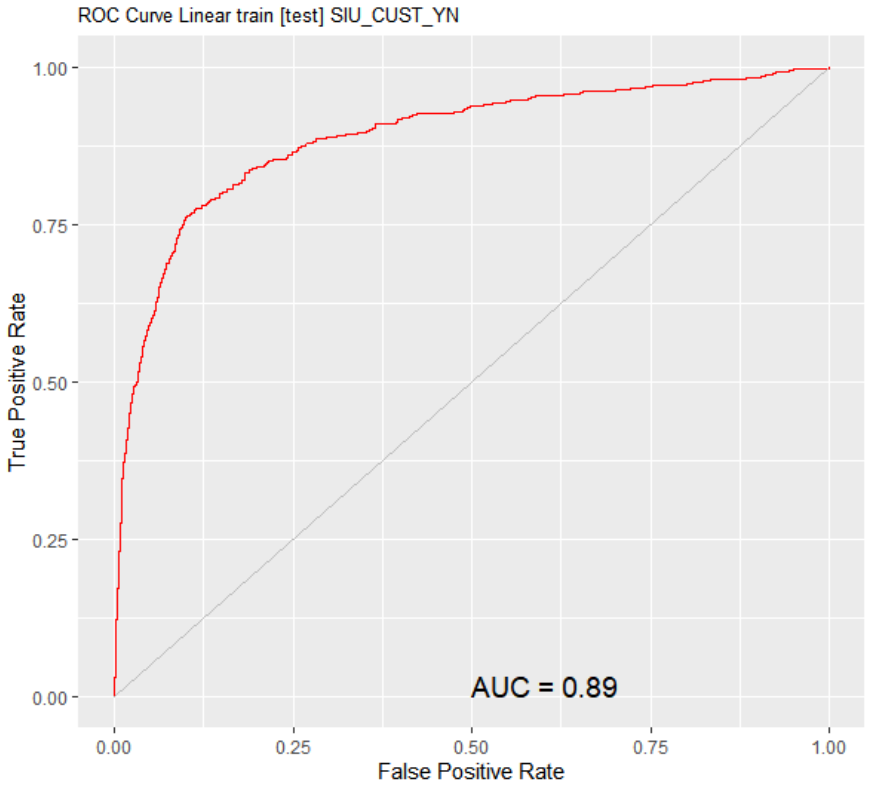
정확도 : 0.9306

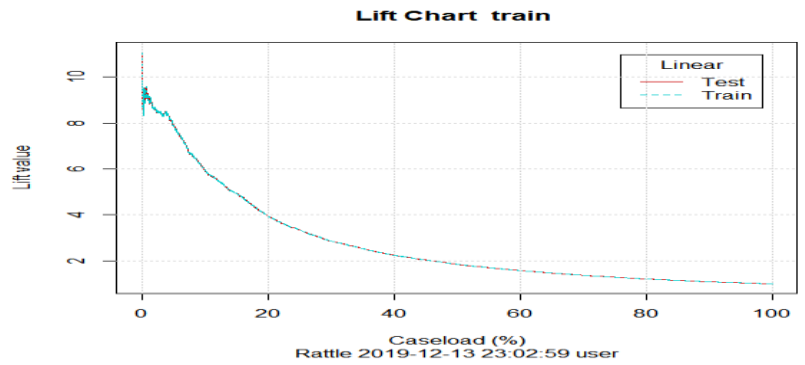
민감도(재현율) : 0.3608

특이도 : 0.9870

정밀도 : 0.7335

F1 - score : 0.4836





Decision Tree

Test Obs = 6,183		Pred	
		No	Yes
True	No	5,541	85
	Yes	357	200

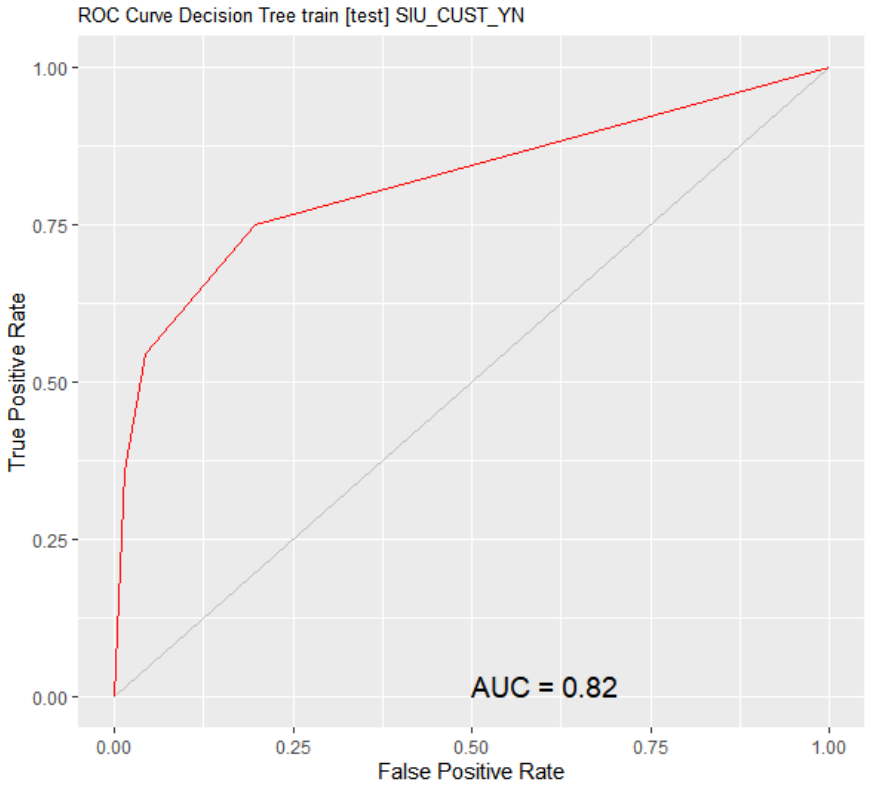
정확도 : 0.9285

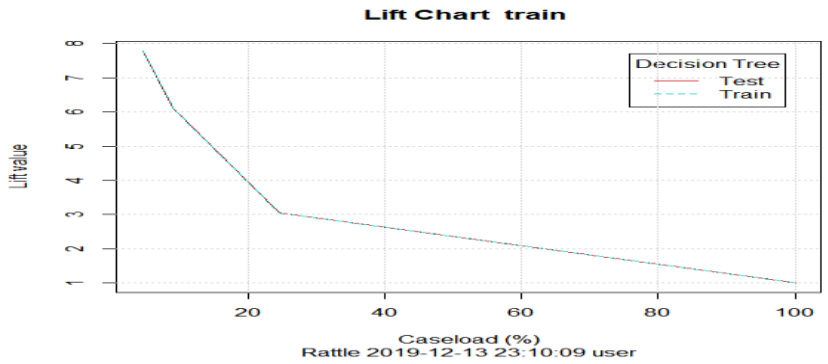
민감도(재현율) : 0.3590

특이도 : 0.9848

정밀도 : 0.7017

F1 - score : 0.4749





Random Forest

Test Obs = 6,183		Pred	
		No	Yes
True	No	5,553	73
	Yes	315	242

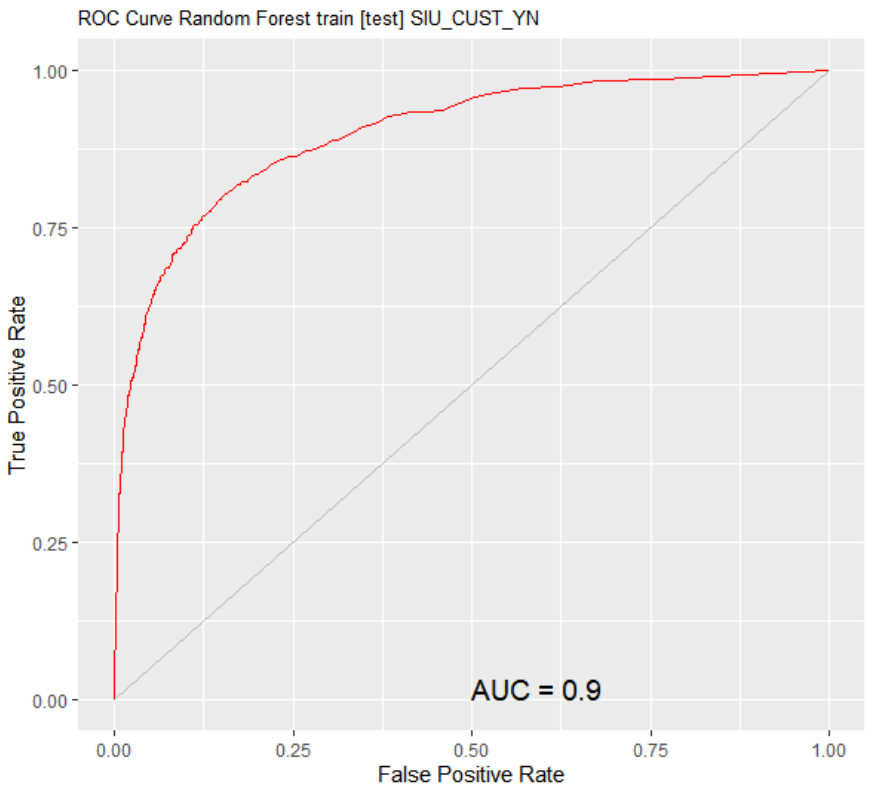
정확도 : 0.9372

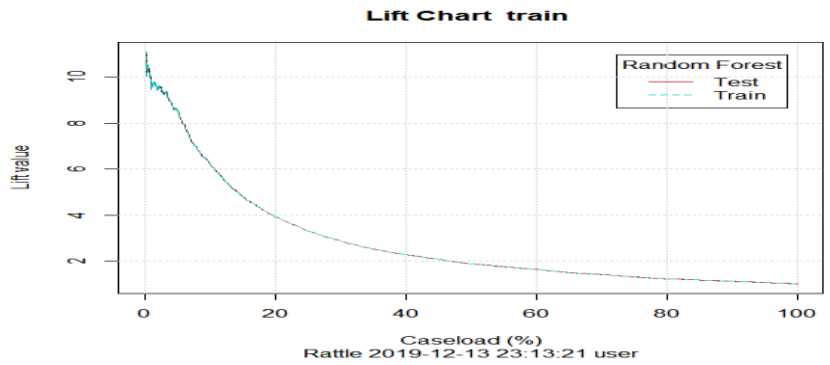
민감도(재현율) : 0.4344

특이도 : 0.9870

정밀도 : 0.7682

F1 - score : 0.5549





Index

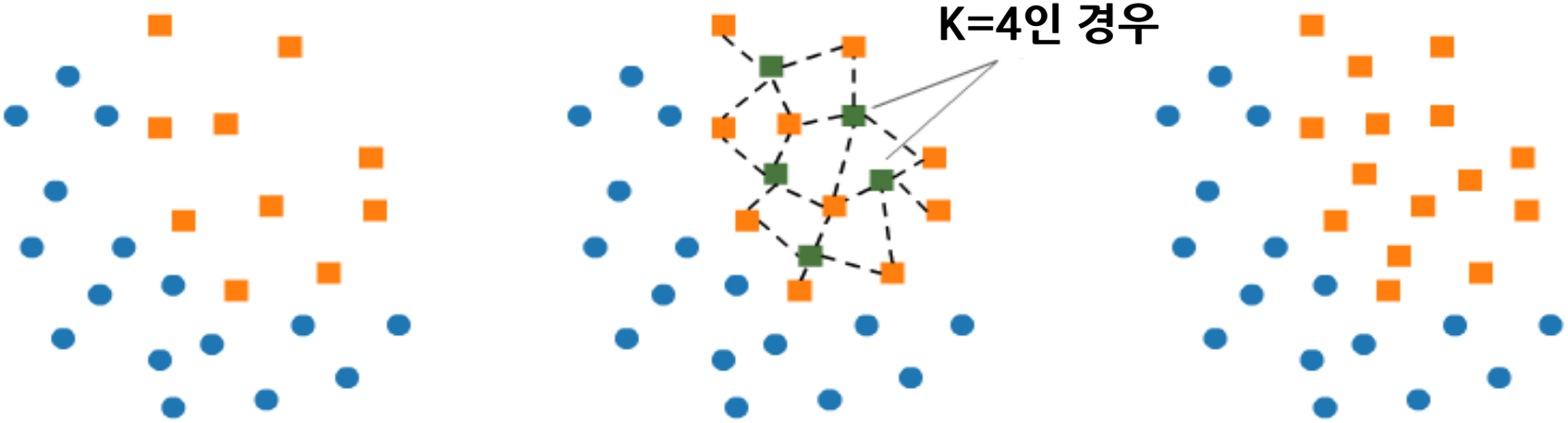
SMOTE

Synthetic Minority Over-sampling Technique의 약자로 비율이 낮은 분류의 데이터를 만들어내는 방법이다.

SMOTE는 먼저 분류 개수가 적은 쪽의 데이터의 샘플을 취한 뒤 이 샘플의 k 최근접 이웃(k nearest neighbor) 을 찾는다.

그리고 현재 샘플과 이들 k개 이웃 간의 차이(difference)를 구하고, 이 차이에 0 ~ 1 사이의 임의의 값을 곱하여 원래 샘플에 더한다.

이렇게 만든 새로운 샘플을 훈련 데이터에 추가한다. 결과적으로 SMOTE는 기존의 샘플을 주변의 이웃을 고려해 약간씩 이동시킨 점들을 추가하는 방식으로 동작한다.



SMOTE

R 프로그래밍의 DMwR 라이브러리의 SMOTE 함수를 이용.

```
DMwR::SMOTE(form,                  # 모델 포뮬러

            data,                  # 포뮬러를 적용할 데이터

            perc.over=200,         # 적은 쪽의 데이터를 얼마나
                                   # 추가로 샘플링해야 하는지

            k=5,                   # 고려할 최근접 이웃의 수

            perc.under=200         # 적은 쪽의 데이터를 추가로 샘플링할 때
                                   # 각 샘플에 대응해서 많은 쪽의 데이터를
                                   # 얼마나 샘플링할지 지정
)
```

SMOTE

```
> data(iris)
> data <- iris[, c(1, 2, 5)]
> data$Species <- factor(ifelse(data$Species == "setosa","rare","common"))
> table(data$Species)
common   rare
   100     50
```

SMOTE 샘플링 진행 후

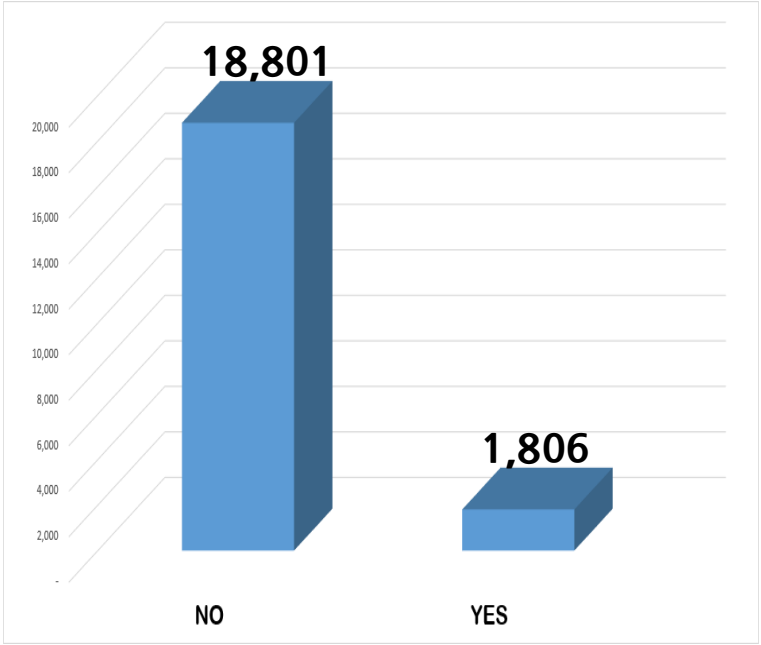
```
> newData <- SMOTE(Species ~ ., data, perc.over = 600, perc.under=100)
> table(newData$Species)
common   rare
   300    350
```

1. Rare 데이터 한 개당 perc.over / 100 (600 / 100 = 6)의 추가 데이터 생성
2. Rare 데이터가 50개에서 300개의 추가 데이터를 더해 350개로 생성
3. 이제, Common 데이터에서 Rare데이터가 추가 생성된 양의 perc.under의 비율만큼 샘플링 진행한다. 즉, Rere가 200개 생성되었으므로, 200 * (100%)인 200개를 역시 common에서 sampling 진행한다.

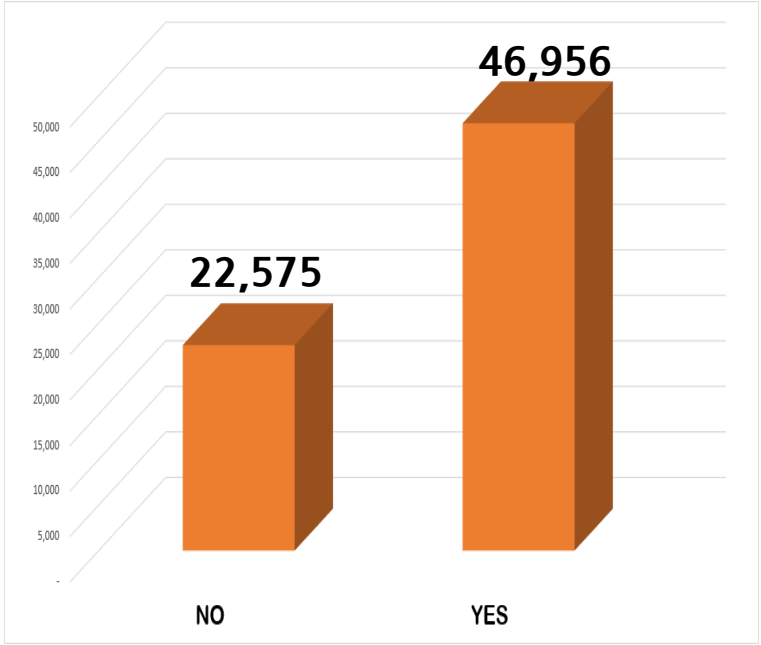
SMOTE

기존의 주어진 데이터는 보험사기자의 수가 1,806명으로 약 8.76%의 Rare Event로 나타났다.

20,607개의 주어진 데이터를 기존의 0.92 : 0.08 의 비율을 역으로 뒤집어 0.33 : 0.67 의 비율로 재 생성



Original Train Data
20,607 obs



SMOTE Train Data
68,531 obs

Random Forest with SMOTE

Test Obs = 20,860		Pred	
		No	Yes
True	No	6,538	277
	Yes	141	13,904

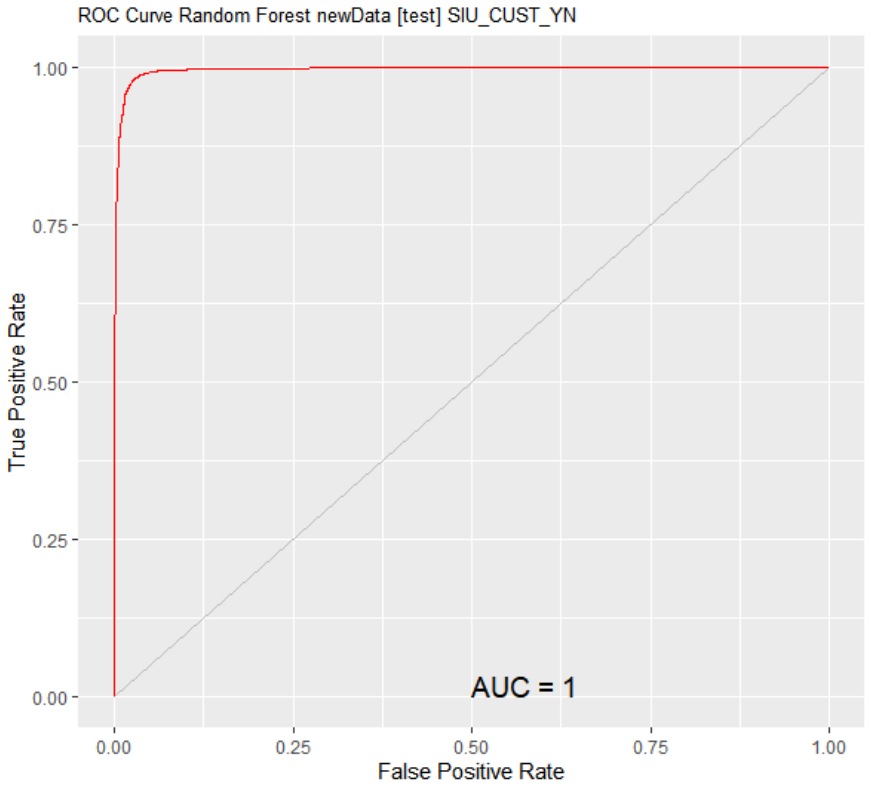
정확도 : 0.9799

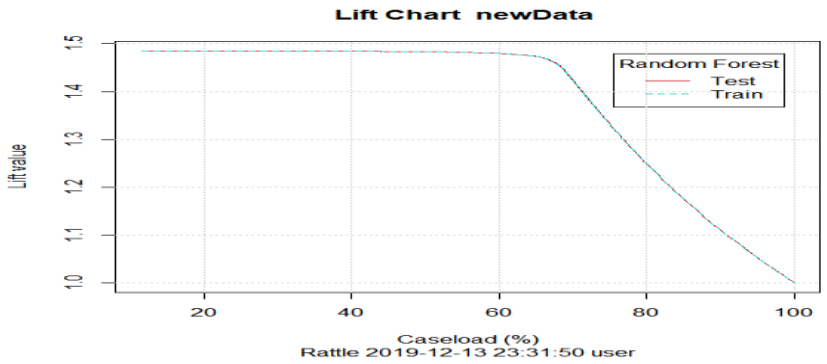
민감도(재현율) : 0.9899

특이도 : 0.9870

정밀도 : 0.9804

F1 - score : 0.9851





Index

Suggestion

일반적으로 보험 사기 적발 모형에서 Logistic Regression이나 Decision Tree같이 단일 모형보다 앙상블 모형인 Random Forest 모형의 예측력이 더 높은 것으로 나타났다.

그러나, 앙상블 모형의 경우 단일 모형에 비해 각 변수들의 설명력이 떨어진다는 단점이 있다. 따라서 구축하고자 하는 모형이 설명력 위주의 모형인지 사기 적중률에 초점을 맞춘 모형인지 올바른 선택이 필요하다.

또한, 모형의 성능을 평가하기 위해 단순히 정확도를 사용하기 보다는 Rare Event의 경우 F1-Score를 사용하는 것이 소수의 보험 사기를 적발하는 모형에 더 적합하다고 판단된다.

Suggestion

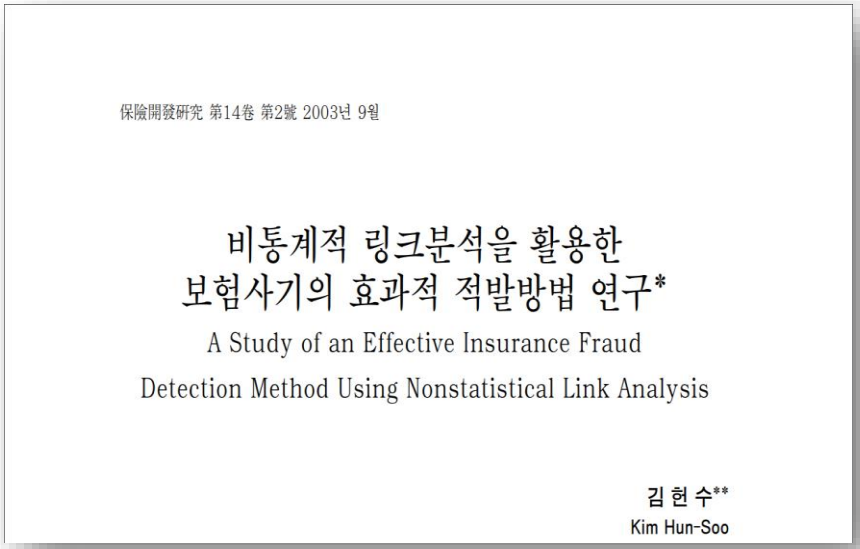
주어진 데이터로 각 개인에 대한 보험 사기자 여부를 판별할 수 있는 모형 뿐만 아니라 각각의 청구 건에 사기로 추정되는 스코어를 부여하는 모형을 만드는 방법도 고려된다.

한 사람이 모든 청구 건수에 대하여 보험 사기를 행하는 일은 매우 드물기 때문에 특정 사건에 대하여 보험 사기를 적발하는 모형도 필요하기 때문이다.

최근 보험사기 추세를 보면 가족단위나 친구 모임,병원이나 보험사직원과 결탁하는 등 집단으로 형성된 보험사기가 급속도로 증가하고 있다.

점차 조직화, 대형화 되어가는 각종 보험사기 범죄를 보다 효율적으로 적발하기 위해서는 조사 대상 건을 선별해주는 스코어링 시스템뿐만 아니라 이들 보험 사기 공모자들간의 배후관계를 파악하고 관련자들을 적발해낼 수 있는 Link Analysis기법을 이용한 시스템을 병행하여 운영할 필요가 있을 것이다.

Link Analysis

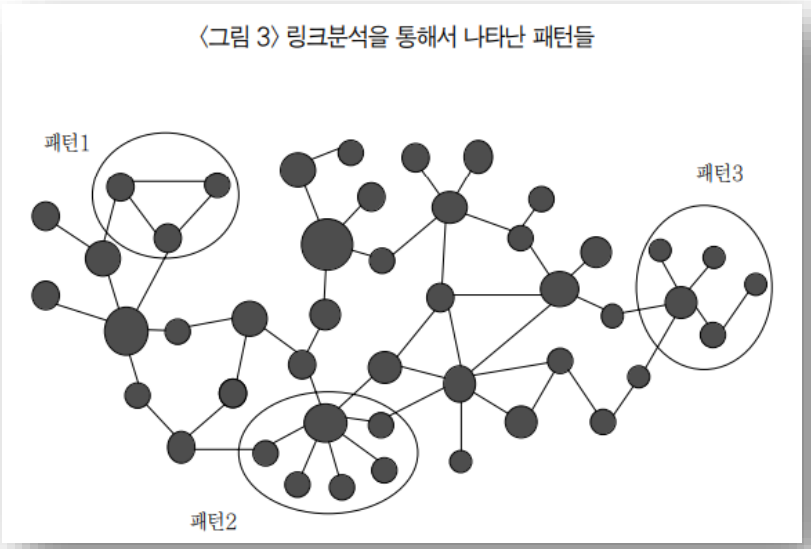


保險開發硏究 第14卷 第2號 2003년 9월

비통계적 링크분석을 활용한
보험사기의 효과적 적발방법 연구*

A Study of an Effective Insurance Fraud
Detection Method Using Nonstatistical Link Analysis

김 헌 수**
Kim Hun-Soo



위 논문에 따르면, 청구 번호, 계약자의 주민번호, 청구자의 주민번호, 사고자 주민번호, 집주소, 전화번호, 자동차 번호, 병원 번호, 정비소 번호 등을 이용하여 고객들에 대하여 링크맵을 구축하고,

다양한 패턴을 발견하여, 각 고객들이 어떤 관계가 있는지, 하나의 보험 사기에 다수의 관련자들이 어떠한 형태로 묶여 있는지 파악하고 보험 사기 혐의자를 탐색하는 분석 방법이다.

THANK YOU